

EXPRESS MAIL LABEL NO.:

(EV 129 136 311 US)

DISTRIBUTION OF DATA VOLUME VIRTUALIZATION

Narasimha R. Valiveti
Ronald S. Karr
Randall K. Shingai
Gopal Sharma

BACKGROUND OF THE INVENTION

[0001] Many business organizations and governmental entities seek fast and inexpensive access to large amounts of data stored in storage area networks. Figure 1 illustrates relevant components of an exemplary data system 10 for storing data. Data system 10 consists of a host node 12 coupled to a storage area network (SAN). The SAN consists of data storage systems 14 and 16 and SAN switch 18.

[0002] Host node 12 takes form in a computer system (e.g., a server) having one or more processors and a memory for storing data or instructions. Host node 12 executes an operating system and a volume manager. Volume managers, such as Volume ManagerTM provided by VERITAS Software Corporation of Mountain View, California, are systems for organizing and managing the distribution of data of a volume across one or more storage devices. Volume and disk management products from other product software companies also provide a system for organizing and managing the distribution of volume data across multiple storage devices.

[0003] Host node 12 may be coupled to one or more client computer systems (not shown). Host node 12 generates input/output (I/O) transactions for reading data from or writing data to the data volume contained in data storage systems 14 and 16. These I/O transactions are transmitted to data storage systems 14 and 16 via SAN switch 18.

[0004] Each of the data storage systems 14 and 16 includes a plurality of storage devices such as hard disks (not shown). For example, data storage system 14 includes three hard disks designated A₁ - A₃, while data storage system 16 includes three hard disks designated

B₁ – B₃. Each of the data storage systems 14 and 16 also include one or more processors for processing I/O transactions received from host node 12 as will be more fully described below.

[0005] As noted above, host node 12 executes a volume manager. The volume manager organizes the hard disks and storage objects (e.g., subdisks, extents, plexes, etc.) created from the hard disks to form a data volume. In organizing these hard disks, the volume manager creates a description of how the data volume is organized or laid out. There are many different ways to organize a data volume from underlying hard disks and storage objects. The layout description relates the storage objects to each other or to the hard disks of the data storage systems.

[0006] Properties of the storage depend on how the volume manager organizes the data volume. In theory, there are a large number of ways to organize the data volume. Popular storage types include spanning storage (using storage from several devices to make a larger volume), striped storage (interleaving block ranges between devices to increase performance), and mirrored storage (storing extra copies of the data to improve reliability). Data system 10 will be described with host node 12 aggregating the hard disks and storage objects of data storage systems 14 and 16 to form mirrored volume storage.

[0007] A mirrored volume replicates data over two or more plexes of the same size. For purposes of explanation, host node 12 aggregates hard disks and storage objects to form a two-way mirrored volume. In this two-way mirror, a logical block number i of a volume maps to the same block number i on each mirrored plex. A two-way mirrored volume corresponds to RAID 1.

[0008] Figure 2 illustrates an exemplary volume layout description for the exemplary two-way mirrored volume stored within systems 14 and 16 of Figure 1. More particularly, Figure 2 consists of two plexes (i.e., plex 1 and plex 2). Plex 1 consists of three subdisks designated subdisk 1 – subdisk 3, while plex 2 consists of three subdisks designated subdisk 4 – subdisk 6. Figure 2 also shows that subdisk 1 – subdisk 3 are allocated from contiguous regions of hard disks A₁ – A₃, respectively, while subdisks 4 – subdisk 6 are allocated from contiguous regions of hard disks B₁ – B₃, respectively. The layout description illustrated in Figure 2, is stored within memory of host node 12. It is noted that the volume manager can modify the layout description as the volume manager modifies the organization of the data

volume. For example, the volume manager may create new, change existing, or destroy storage objects of the volume.

[0009] Host node 12 uses volume layout description for many purposes. For example, host node uses the volume layout description illustrated in Figure 2 when writing data to or reading data from the data volume. To illustrate, when host node 12 seeks to write data D to block x of the mirrored data volume example, host node 12 accesses the volume layout description shown in Figure 2 to determine the location of the plexes to be updated with data D. In the illustrated example, the volume layout description indicates that data D is to be written to plex 1 and plex 2 aggregated from hard disks located in data storage systems 14 and 16, respectively.

[0010] Because the plexes are located in different data storage systems, host node 12 must generate and transmit separate I/O transactions to write data D to the data volume. More particularly, host node 12 generates first and second I/O transactions for writing data D to block x in plex 1 and plex 2, respectively. The first and second I/O transactions are sent to data storage systems 14 and 16, respectively. Data storage systems 14 and 16 process the first and second I/O transactions, respectively, and write data D to respective hard disks. A high frequency of I/O transactions transmitted between host node 12 and data storage systems 14 and 16 may burden the data system 10.

SUMMARY OF THE INVENTION

[0011] After a first device in a network creates or modifies a description of a layout for a data volume, the first device transmits separate copies of the data volume layout description to a pair of second devices, respectively, for storage in respective memories thereof. The first device may be a host node, and the pair of second devices may be first and second data storage systems. The first device and the pair of second devices are configured so that I/O transactions are transmitted between the first device and either of the pair of second devices. In other words, the first device is contained in a network layer that is different from the network layer that contains the pair of second devices.

BRIEF DESCRIPTION OF THE DRAWINGS

[0012] The present invention may be better understood, and its numerous objects, features, and advantages made apparent to those skilled in the art by referencing the accompanying drawings.

Fig. 1 is a block diagram of a data system;

Fig. 2 illustrates an exemplary volume layout description employed in the data system of Fig. 1;

Fig. 3, is a block diagram of a data system employing one embodiment of the present invention, and;

Fig. 4 illustrates an exemplary volume layout description employed in the data system of Fig. 3.

[0013] The use of the same reference symbols in different drawings indicates similar or identical items.

DETAILED DESCRIPTION

[0014] Figure 3 illustrates an exemplary data system 20 in which one embodiment of the present invention may be employed. The data system includes a first device 22 coupled to a pair of second devices 24 and 26 via a third device 28. The pair of second devices 24 and 26 are contained in a layer that is different than the layer that contains the first device 22 or the layer that contains the third device 28. I/O transactions are transmitted between devices in different layers of the data system 20. Thus, I/O transactions are transmitted between the first device 22 and device 24 via the third device 28, and I/O transactions are transmitted between the first device 22 and device 26 via device 28. For purposes of explanation, first device 22 takes form in a host node, the pair of second devices 24 and 26 take form in data storage systems while the third device 28 takes form in a SAN switch. The present invention will be described with reference to data system 20 containing a host node 22, a SAN switch 28, and two data storage systems 24 and 26, it being understood that the present invention should not be limited thereto. For example, the present invention could be used in a system in which Fiber Channel connected, IP-connected, and/or Infiniband-connected storage is used. Also, the term coupled should not be limited to a direct

connection between two devices. Rather, the term coupled includes an indirect connection between two devices (e.g., host node 22 and data storage system 24) via a third device (i.e., SAN switch 28) so that the two devices can communicate via the third device.

[0015] Host node 22 may take form in a computer system (e.g., a server) having one or more processors and memory for storing data or computer executable instructions. Host node 22 may execute an operating system and a volume manager. It is noted that the volume manager may execute on a device other than host node 22. As will be more fully described below, the volume manager or another unit within host node 22, generates a description of the layout of the data volume distributed across devices 24 and 26. However, it should be noted that generation of the data volume description layout need not be generated at host node 22. Rather, the data volume layout description could be generated at a device such as node 30 coupled to host node 22. Moreover, the volume layout description could be generated in one or the storage systems 24 or 26, or in SAN switch 28. The device that generates the volume layout description distributes copies to one or more devices in system 20. For example, if the volume layout description is generated at node 30, node 30 could provide a volume layout description copy to host node 22 (or another node which is in the same layer as host node 22) and one or more devices in separate layers, e.g., copies provided data storage systems 24 and 26, or a copy provided to just SAN switch 28. The remaining description will presume that the data volume description layout is generated at the host node 22, and that host node 22 distributes copies of the volume layout description to one or more other devices in system 20, it being understood that the present invention should not be limited thereto.

[0016] Host node 22 generates I/O transactions to read data from or write data to the data volume stored in data system 20. For purposes of explanation, data system 20 will be described as storing just one two-way mirrored data volume, it being understood the present invention can be employed with data system 20 storing several data volumes. Moreover, it should be noted that the present invention should not be limited to data system 20 storing a mirrored volume. The present invention may be employed in a data system 20 employing parity or other forms of calculated redundancy, striping, and aggregation, along with features such as snapshots, replication, and online reorganization.

[0017] Each of data storage systems 24 and 26 may take any one of many different forms. For example, data storage systems 24 and 26 may take form in intelligent disk arrays, block server appliances, or combinations thereof. Each of the data storage systems 24 and 26 may include a plurality of storage devices (not shown) for storing volume data. Each of these storage devices may take form in one or more dynamic or static random access memories, one or more magnetic or optical data storage disks, or combinations thereof. Data storage system 24 will be described as having three storage devices designated HD₁ - HD₃, while data storage system 26 will be described as having three storage devices designated HD₄ - HD₆. For purposes of explanation, storage devices HD₁ - HD₆ take form in hard disks, it being understood that storage devices should not be limited thereto. The storage devices of storage systems 24 and 26 could take form in any hardware, software, or combination of hardware and software in which data may be persistently stored and accessed.

[0018] Data storage systems 24 and 26 may include one or more processors and memory for storing computer executable instructions. Data storage systems 24 and 26 are capable of processing I/O write transactions received from host node 22 as will be more fully described below. Data storage system 24 can write data to one or more hard disks HD₁ - HD₃ in response to data storage system 24 receiving and processing an I/O write transaction, and data storage system 26 may write data to one or more hard disks HD₄ - HD₆ in response to data storage system 26 receiving and processing an I/O write transaction.

[0019] As noted above, host node 22 executes a volume manager. The volume manager 22 organizes the hard disks and storage objects (e.g., subdisks, extents, plexes, etc.) created from the hard disks of system 20 to form a data volume. In organizing these hard disks and storage objects, the volume manager creates a description of how the data volume is organized or laid out. There are many different ways the volume manager can organize a data volume from underlying hard disks and storage objects. For purposes of explanation only, the volume manager organizes the hard disks and storage objects to form a two-way mirrored volume V.

[0020] Figure 4 illustrates an exemplary volume layout description created by the volume manager of host 22 for the two-way mirrored volume. The layout description of Figure 4 consists of plex 1 and plex 2. A logical block number i of a volume V maps to the

same block number i on each mirrored plex. Plex 1 consists of three subdisks designated subdisk 1 – subdisk 3, while plex 2 consists of three subdisks designated subdisk 4 – subdisk 6. It is noted that mirrored plexes need not contain the same number of subdisks. Figure 4 also shows that subdisk 1 – subdisk 3 is allocated from contiguous regions of hard disks HD₁ – HD₃, respectively, while subdisks 4 – subdisk 6 are allocated from contiguous regions of hard disks HD₄ – HD₆, respectively.

[0021] After the volume layout description of Figure 4 is first created, host node 22 can transmit a copy thereof to any one or more of the devices in system 20 including data storage system 24, data storage system 26, and/or SAN switch 28. In another embodiment, the copies of the volume layout description transmitted to the various devices in system 20 need not be identical to each other. Host node 22 could transmit copies of the volume layout description that are tailored to the operating characteristics of the devices that receive them.

[0022] The host node 22 may modify the volume layout description to account for changes in the corresponding volume. More particularly, the volume layout description is modified each time the volume manager of host node 22 creates new, changes existing, or destroys storage objects of volume V. It is important that distributed copies of the volume layout description are maintained consistent with each other. To achieve consistency when modifications are made to the volume layout description, host node 22 transmits copies of the modified volume layout description to each device (e.g., data storage systems 24 and 26) that received a prior copy. Alternatively, host node 22 transmits information that enables devices (e.g., data storage systems 24 and 26) to modify their copies of the volume layout description.

[0023] Once data storage systems 24 and 26 have a copy of volume layout description from host node 22, data storage systems 24 and 26 are capable of performing new operations. To illustrate, host node 22 may prepare and transmit an I/O transaction to write data D to, for example, data storage system 24. It is noted that in one embodiment, host node 22 may alternate between data storage system 24 and data storage system 26 as the destination for subsequent I/O write data transactions in an attempt to load balance the I/O write transactions within system 20. It should be made clear that there are many uses of the distributed volume layout description within a given system, and that the following

description represents just one use. Further, the distributed volume layout description can be used for several distinct purposes within a system.

[0024] In response to receiving the I/O write transaction from host node 22, data storage system 24 accesses its local copy of volume layout description to identify the plexes where data D is to be written. In the illustrated example, data storage system 24 determines that data is to be written to each mirror (i.e., plex 1 and plex 2) of the mirrored volume V. Data storage system 24 recognizes from volume layout description that plex 1 is an aggregation of subdisks which have been allocated from the hard disks HD₁ – HD₃ of data storage system 24, and that plex 2 is an aggregation of subdisks which have been allocated from the hard disks HD₄ – HD₆ of data storage system 26.

[0025] Data storage system 24 writes data D to one or more of hard disks HD₁ – HD₃ after accessing its copy of the most current volume layout description. Control information may cause data storage system 24 to forward the I/O transaction to data storage system 26 in response to determining from the volume layout description that plex 2 is contained within data storage system 26. The I/O transaction may be forwarded to data storage system 26 with some modification. For example, the I/O transaction may be forwarded with an instruction that data storage system 26 should not send the I/O transaction back to data storage system 24. It is noted that in an embodiment where data of volume V is distributed over more than two data storage systems of data system 20, data storage system 24 may forward the write I/O transaction to all data storage systems (other than data storage system 24), or data storage system 24 may forward the I/O transaction in multicast fashion to only those data storage systems that contain plexes where data D is to be written according to the volume layout description. Each data storage system that receives the I/O transaction could access its copy of the volume layout description to determine whether data D is to be written to one or more of its storage devices.

[0026] Data storage system 26, in response to receiving the I/O transaction from data storage system 24, may access its local copy of volume layout description and determine that data D is to be written to plex 1 and plex 2. Since data storage system 26 recognizes that it stores plex 2, data storage system 26 writes data D to one or more of hard disks HD₄ – HD₆. After data storage 26 writes data D, data storage system 26 optionally transmits a message to data storage system 24 indicating that plex 2 has been updated with data D. Data storage system 24, in turn, may optionally transmit a message to host node 22

indicating that plex 1 and/or plex 2 have been updated with the new data D in response to receiving the update message from data storage system 26.

[0027] As noted above, distributed copies of the data volume layout description should be consistent with each other. In one of the examples, when host node 22 modifies its copy of the data volume layout description, copies of the modified volume layout description are provided to each device that received a prior version of the volume layout description. Each device subsequently updates its copy of the volume layout description. It may be necessary to delay, for example, host node 22's transmission of new I/O transactions until all devices update their copies of the volume layout description and consistency is again obtained between the distributed volume layout descriptions. The delay in transmission of new I/O transactions may begin with the first phase of a two phase commit. The first phase pauses I/O processing at the host node 22, and the second phase unblocks I/O processing at the host node 22 when modifications to the distributed copies of the volume layout description have been committed. In this fashion, data coherency is maintained in the data volume before and after modification of the volume layout description.

[0028] Although the present invention has been described in connection with several embodiments, the invention is not intended to be limited to the specific forms set forth herein. On the contrary, it is intended to cover such alternatives, modifications, and equivalents as can be reasonably included within the scope of the invention as defined by the appended claims.